**Multivariate data analysis in gene therapy process development**

Joe Emerson1, Mark J. Willis1, Jarka Glassey1

*1 School of Engineering, Newcastle University, Merz Court, Newcastle upon Tyne, NE1 7RU*

*\*Corresponding author: jtemerson@newcastle.ac.uk*

**Highlights**

* Multivariate techniques were applied to historical process data from vector production and cell processing
* Principal component analysis was used to assess batch-to-batch process variability
* Partial least squares enabled identification of critical process variables and modelled their relationship with critical quality attributes of the product

**1. Introduction**

Recent developments to vector safety and efficacy have brought about significant renewed interest in gene therapy and evidence of successful clinical trials has been mounting. However, the production processes are at an early stage of development, and process complexity and variability in vector/cell manufacturing processes, combined with the data-lean features of the early development stages in cell-gene therapy are significant contributing delaying factors in the new product development. In this work advanced modelling approaches are applied to process information from vector and cell expansion to bring benefits in process understanding and contribute to more rapid development of new products.

**2. Methods**

The nature of the data set is typical of bioprocesses, with a high number of inputs (process parameters) and outputs (performance parameters) [1]. In addition, the number of batches is low in many cases due to the treatments being personalized or targeted at rare diseases. The combination of many variables with a low number of repeats is a challenge for modelling practice. In this work, standard techniques were employed though it was necessary to modify the approach to modelling in order to cope with the challenges of the data set. For example, the use of sparse principal component analysis (PCA) was explored as an alternative to standard PCA in order to ease the interpretation of loadings with a high number of variables. Partial least squares (PLS) is an extension of multivariate linear regression, used to model the relationship between a set of independent variables and output variables of interest. PLS models were developed to link process variables to critical quality attributes of the product, critical process parameters were identified, and validation techniques were implemented to assess the predictive capability of the models. The validated models were interpreted to gain valuable process insights.

**3. Results and discussion**

Demonstrative PCA results are presented here along with a discussion, full results from PCA, sparse PCA and PLS modelling will be presented in the paper. PCA was applied to the production process variables and figure 1 shows the scores for the first three principal components. Figure 1 demonstrates the usefulness of PCA for highlighting and visualizing the differences in operating conditions that occur batch-to-batch. Batches that are close together have similar operating conditions whilst the opposite is true for batches that are further apart. Colouring the batches by the viral titre clearly shows that the operating conditions exhibited in batches 9 to 13 were favourable for a high titre.



**Figure 1.** PCA applied to process variables: scores for principal components 1 to 3

To understand the driving forces behind the differences in batch scores, it is necessary to analyse the PCA loadings vector, in this way it is possible to identify variables with high variability and highlight the variables that were changed batch-to-batch and cluster-to-cluster. PLS was used model the relationship between process variables and the viral titre. Using variable selection techniques, it was possible to identify critical process parameters i.e. those influencing the viral titre and responsible for the high titre of batches 9 to 13 on figure 1. The PLS models were cross validated, and models with high predictive capability were selected. The sign on the standardized regression coefficients provided information on whether the variables positively or negatively influenced the viral titre, and the magnitude of coefficients indicated the relative importance of their effect.

**4. Conclusions**

Multivariate data analysis is a powerful tool for learning from historical process data; which may be particularly useful in cell gene therapy where complex biology and a large number of process parameters interacting, leads to a lack of process understanding. PCA is effective at identifying areas of the process with high variability and characterising the changes that were made during process development. The PLS models developed provided useful insights for process development, including identifying critical process parameters and indicating how these variables may be manipulated to increase the viral titre.

**References**

1. A.S. Rathore, S. Mittal, M. Pathak, A. Arora, *Biotechnology Progress*, 2014, *30*(4), pp 967–973.